**UNITED STATES DEPARTMENT OF COMMERCE**
**Bureau of the Census**
Washington, DC 20233-0001

February 13, 2001

**DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter DT-6**

| | |
|---|---|
| **Memorandum For:** | **Magda Ramos**<br>**Chief, Coverage Measurement Operations Staff** |
| **From:** | **Danny Childers** DRC<br>**Leader, A.C.E. Design Team** |
| **Subject:** | **Algorithm for Determining Outlier Clusters in the 2000 A.C.E.** |
| **Prepared By:** | **John Jones** |

## 1. Introduction

During Housing Unit Matching, BFU Person Matching, and AFU Person Matching the Design Team will manually select A.C.E. clusters for Outlier Review. At the end of AFU Batch Processing in Person Matching, clusters for Outlier Review will be chosen automatically, by an algorithm that assigns a score to each cluster. The score for each cluster will reflect the weighted number of P-sample nonmatches, the weighted number of P-sample and E-sample unresolved persons, and the weighted number of erroneous enumerations. Clusters achieving a score above a cutoff will be automatically chosen for outlier review.

## 2. The Algorithm for Person Matching

For each cluster we calculate the score as:

$$\text{SCORE}= \left( \sum_{PSN} PSW + \sum_{PSU} PSW + \sum_{EE} ESW + \sum_{ESU} ESW \right) \Big/ \sqrt{T}$$

Where:

PSN are the P-sample nonmatches, i.e. those with P-sample match codes NP, NC, NU, NR.

PSU are the P-sample unresolveds, i.e. those with P-sample match codes KI, KP and all persons with match codes P and MU.

EE are the E-sample erroneous enumerations, i.e. those with E-sample enumeration codes EE, GE, KE, DE, FE, MN.

ESU are the E-sample unresolveds, i.e. those with E-sample enumeration codes UE, GU.

T is the unweighted sum of those P-sample persons with match codes M, MR, MU, NP, NC, NR, NU, P, KI, KP, and those E-sample persons with match codes CE, GE, EE, FE, DE, KE, UE, GU, MN. (Descriptions of all codes are attached)

PSW is the final P-sample cluster weight
ESW is the final E-sample cluster weight

Both PSW and ESW are weights whose source is the Sample Design File maintained by the Sample Design Team. PSW is the trimmed weight for P-sample housing units in a cluster and it is denoted TRIMWTP in the Sample Design File. ESW is the trimmed weight for E-sample housing units in a cluster. There are two variables in the Sample Design File corresponding to this trimmed weight; TRIMWTE1 for housing units with ESPS code 1, and TRIMWTE2 for housing units with ESPS code 2. (see DSSD Census 2000 Procedures and Operations Memorandum Series R-4, Appendix B for the Sample Design File Layout). The variables TRIMWTP, TRIMWTE1, and TRIMWTE2 are also on the Cluster Control File for Clerical Person Matching (Layout name: ACELAY:PERMARCS_ACCT). The corresponding names are PWGT, E1WGT, and E2WGT. Set ESW=TRIMWTE1 for those persons in housing units with ESPS code 1 and set ESW=TRIMWTE2 for those persons in housing units with ESPS code 2. The ESPS code is discussed in Section 4.3 of the Design Document (DSSD Census Procedures and Operations Memorandum Series, Chapter S-DT-1 by Danny Childers) where it is called E-sample probability code. This code identifies the proper weight to use when the probability of selection is not the same for all E-sample people in the cluster.

The threshhold value of the score is 2600. However, this threshhold needs to be adjustable. Clusters whose score exceeds this threshhold will be chosen for Outlier Review.

3. **Priority Scheme**

Analysts will review clusters that qualify for Outlier Review based upon the score. The clusters with the highest scores will be given priority.

## 4.14 Final P-sample Person Match Codes

The probability of being matched is estimated for the P-sample people with unresolved match status.

### 4.14.1 Matched

M = The P-sample and the census people were matched.

MR = The P-sample follow-up interview determined that the matched person with unresolved residence status is a resident. The person is a P-sample person and is matched to a E-sample person.

MU = The A. C. E. person follow-up interview obtained no useful information to resolve the residence status for the matched person who had a residence status of unresolved before follow-up. The P-sample person's residence status is unresolved and the E-sample person's enumeration status is unresolved.

### 4.14.2 Not Matched

NP = The P-sample person is not matched to a census person. There was no follow-up for the whole household nonmatches from person interviews with household members and the whole household nonmatches were not conflicting household nonmatches.

NC = The P-sample nonmatch was found on the census roster. This person in a partial nonmatch household was not matched to the census because only name was collected in the census for this person in a large household and the census person was not data defined. No follow-up interview is necessary.

NR = The P-sample person is identified as a resident in the block cluster on census day during the A. C. E. person follow-up interview. The P-sample person is missed in the census.

NU = Not enough information is collected during the A. C. E. person follow-up interview to identify the P-sample person as a resident or nonresident in the block cluster. The residence status for the P-sample person is unresolved. This code is also used when the a P-sample person is followed up to collect geographic information and that information is not collected.

### 4.14.3 Unresolved

P = There is not enough information collected to determine if the possible match is a

match or not. The match status of the P-sample person and the E-sample person is unresolved.

KI = Match not attempted for the P-sample person because the person has insufficient information for matching and follow-up. The name is blank or incomplete or the name is complete but the person has only one characteristic. This is a computer assigned code and these people are suppressed from view by the matchers.

KP = Match not attempted for the P-sample person, because (1) the name is incomplete, such as "Mr. Jones", or (2) the name is not a valid name, such as "White Female" or "Donald Duck". This is a clerically assigned code.

### 4.15 E-sample Person Enumeration Codes

The probability of being correctly enumerated is estimated for the E-sample people with unresolved enumeration status.

#### 4.15.1 Correctly Enumerated

M = The P-sample and E-sample people were matched. The E-sample person is correctly enumerated.

CE = The E-sample nonmatch is identified as correctly enumerated during the A. C. E. person follow-up interview.

MR = The A. C. E. person follow-up interview determined that the matched person with unresolved residence status is a resident. The person is a P-sample person and is matched to an E-sample person.

#### 4.15.2 Erroneously Enumerated[1]

GE = The E-sample person is erroneously enumerated in this block cluster, because the census housing unit is a geocoding error (i.e., counted in the block cluster in error). The E-sample person should have been enumerated elsewhere in the census.

EE = The E-sample nonmatch is identified during the person follow-up interview as

---

[1] The E-sample people who are duplicated with census people with E-sample indicator of 2 are not full erroneous enumerations. If the E-sample person with an E-sample indicator of 1 is duplicated once with a census person with an E-sample indicator of 2, the E-sample person is given one half of an erroneous enumeration. If the E-sample person is duplicated twice with non E-sample people in the cluster, the E-sample person is given two thirds of an erroneous enumeration. The formula is the number of times duplicated is d and the proportion of erroneous enumeration for the E-sample person is d/(d+1). This assumes the E-sample person has been coded as correctly enumerated. If the E-sample person is coded unresolved, the final probability of erroneous enumeration includes an imputation for unresolved enumeration status. If the E-sample person is assigned a match code that indicates erroneous enumeration, the number of times that the E-sample person is duplicated with non E-sample people is irrelevant and ignored. A person can not have a probability of erroneous enumeration that is larger than 100 percent.

erroneously enumerated.

FE = The E-sample nonmatch is determined to be fictitious in this block cluster during the follow-up interview. The person may have existed, but should not have been enumerated in the census within this block cluster. The E-sample person is erroneously enumerated in the census in this block cluster.

DE = The E-sample person is a duplicate of another E-sample person. The code is also used when the E-sample person is a duplicate of a census person in a surrounding block. The people in the E-sample housing unit are erroneously enumerated because they were counted accurately in the surrounding block and duplicated in the sample block cluster.

MN = The A. C. E. person follow-up interview determined that the matched person with unresolved residence status is not a resident in this housing unit or in this block cluster. The person is no longer in the list of P-sample people and the E-sample person is an erroneous enumeration.

KE = Match not attempted for the E-sample person. The name is blank or incomplete or the name is complete but the person has only one characteristic, which is assigned by computer. The name is incomplete or not a valid name, such as "Child Jones", or "Mickey Mouse", which is assigned clerically.[2]

### 4.15.3 Unresolved

UE = Not enough information is collected during the A. C. E. person follow-up interview to identify the E-sample person as correctly or erroneously enumerated in the E-sample. The enumeration status for the E-sample person is unresolved.[3] This code is also used when the a P-sample person is followed up to collect geographic information and that information is not collected.

MU = The A. C. E. person follow-up interview obtained no useful information to resolve the unresolved residence status for the matched person. The P-sample person's residence status is unresolved and the E-sample person's enumeration status is unresolved.

P = There is not enough information collected to determine if the possible match is a match or not. The status of the P-sample person and the E-sample person is unresolved.

GU = The geographic work for the targeted extended search is unresolved. The code has the same definition in both the before and after follow-up matching. The difference is in after follow-up, the code is only used in the list/enumerate

---

[2]There are two types of insufficient information. The insufficient information for matching and follow-up are coded as KE by the computer or clerically and are treated as erroneous enumerations. The insufficient information in the dual system estimator are whole person imputations and are subtracted from the census count of persons.

[3] The UE code is also used when the person did not live at the sample address on Census Day and the Census Day address is not complete enough to determine if the census day address is in the sample block cluster.

clusters. The field work for the targeted extended search was not done or the block number on the form was not in the surrounding blocks, in the block cluster, or on the map. It is not clear where the housing unit is located.